Customer Churn Prediction Report

# 1. Introduction

This report summarizes the customer churn prediction task using a dataset containing 1,000 customers. We explored various modeling techniques to predict churn (20% of customers churned).

# 2. Dataset Overview

The dataset included demographic, transaction, customer service, online activity, and churn status information. The final dataset was fully numeric after one-hot encoding categorical features and engineering a 'DaysSinceLastLogin' feature.

# 3. Models and Performance

## 3.1 Logistic Regression

Accuracy: 51.5%  
ROC-AUC: 0.505  
Churner Recall: 54% (22/41 churners identified)  
Churner Precision: 22%  
Confusion Matrix:  
[[81, 78],  
 [19, 22]]

## 3.2 Random Forest

Accuracy: 78.5%  
ROC-AUC: 0.488  
Churner Recall: 0% (model ignored churners)  
Confusion Matrix:  
[[157, 2],  
 [41, 0]]

## 3.3 Balanced Random Forest (Tuned)

Accuracy: 52%  
ROC-AUC: 0.491  
Churner Recall: 44% (18/41 churners identified)  
Churner Precision: 20%  
Confusion Matrix:  
[[86, 73],  
 [23, 18]]

## 3.4 XGBoost

Accuracy: 57%  
ROC-AUC: 0.484  
Churner Recall: 24% (10/41 churners identified)  
Churner Precision: 16%  
Confusion Matrix:  
[[105, 54],  
 [31, 10]]

# 4. Key Takeaways

- Models consistently struggled with the imbalance and weak signals.  
- Balanced Random Forest showed the best recall for churners (44%) but at a cost of false positives.  
- XGBoost and Logistic Regression had modest performance with slightly better balance between classes.  
- ROC-AUC values below 0.5 suggest very limited predictive separability with current features.

# 5. Final Notes

Further improvement would require:  
- Feature engineering and data enrichment (e.g., more customer behavior data).  
- Possible use of domain-specific insights to improve churn prediction.  
- Exploring ensemble stacking techniques or boosting methods with advanced tuning.